Fairness guarantee in multi-class classification

Christophe Denis

Joint work with:
R. Elie, M. Hebiri, and F. Hu

LAMA, Université Gustave Eiffel

31/09/2022

Journées MAS, Rouen

Multi-class classification: *setting*

Framework

- observation $X \in \mathcal{X}$ and $Y \in \mathcal{Y} = \{1, \ldots, K\}$
- classifier $g : \mathcal{X} \to \mathcal{Y}$
- misclassification risk $R(g) = \mathbb{P}(g(X) \neq Y)$

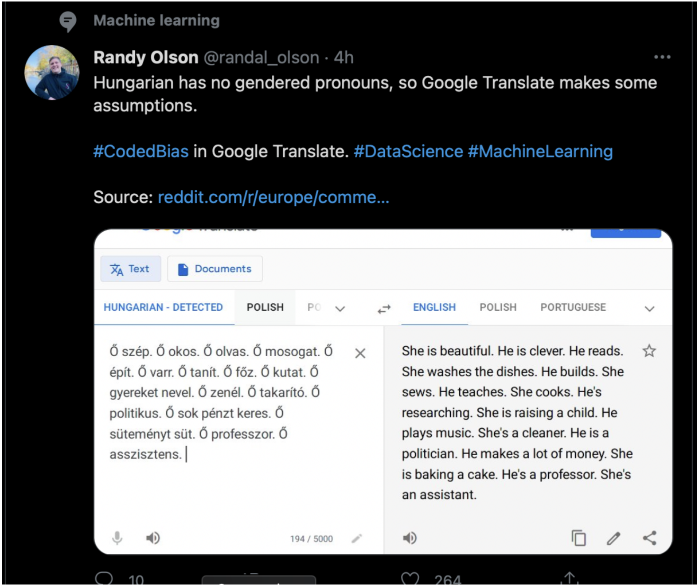
Optimal rule

- conditional probabilities $p_k(X) = \mathbb{P}(Y = k \big| X)$
- Bayes classifier $g^*(\cdot) \in \arg\max_{k \in \mathcal{Y}} p_k(\cdot)$
- oracle risk $R^* = R(g^*) = \min_g R(g)$

Goal

- learning sample $(X_i, Y_i)_{1 \leq i \leq n}$ and new observation X_{n+1}
- empirical classification rule $\widehat{g}$ based on the learning sample
- $\hat{g}(X_{n+1})$ prediction of the associated label

Fairness: *Motivating example*



Algorithmic Fairness

Motivation

- mitigate the bias contained in historical data
- reduce influence of a sensitive attributes in prediction
- lot of attention in recent years Calders *et al.* (2009), Zemel *et al.* (2013), Zafar *et al.*, Donini *et al.* (2018), Agarwal *et al* (2018), Barocas *et al.* (2019), ...

Application

- social sicences
- insurance
- artificial intelligence, ...

Fairness in multi-class classification: *group fairness approach*

Framework

- Obervation (X, S) and $Y \in \mathcal{Y}$,
- $S \in \{-1, 1\}$ sensitive attribute
- classifier $g \to$ prediction $g(X, S)$

Definition of fairness

- Demographic parity (DP), for each $k \in \mathcal{Y}$

$$\mathbb{P}\left(g(X, S) = k | S = 1\right) = \mathbb{P}\left(g(X, S) = k | S = -1\right)$$

- Equalized odds, for each $k \in \mathcal{Y}$

$$\mathbb{P}\left(g(X, S) = k | S = 1, Y = k\right) = \mathbb{P}\left(g(X, S) = k | S = -1, Y = k\right)$$

Multi-class classification under DP constraint

Problem

- $\pi_s = \mathbb{P}(S = s) > 0$, et $p_k(X, S) = \mathbb{P}(Y = k|X, S)$
- $g^* \in \arg\min_g \{\mathbb{P}(g(X, S) \neq Y), g$ satisfies DP$\}$
- lagrangian associated to the minimization problem

$$\mathcal{R}_\lambda(g) = \mathbb{P}\left(g(X, S) \neq Y\right) + \sum_{k=1}^{K} \lambda_k \sum_{s \in \mathcal{S}} s\mathbb{P}(g(X, S) = k|S = s)$$

Continuity assumption

- $t \mapsto \mathbb{P}(p_k(X, S) - p_j(X, S) \leq t|S = s)$ is continuous

Optimal fair classifier

Optimal predictor

- the optimal fair classifier g^* can be characterized as

$$g^*(x, s) \in \arg\max_k \left(p_k(x, s) - \frac{s}{\pi_s} \lambda_k^* \right)$$

- λ_k^* are lagrange multiplier defined as

$$\lambda^* \in \arg \min_{\lambda \in \mathbb{R}^K} \sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s} \left[p_k(X, s) - \frac{s}{\pi_s} \lambda_k^* \right]$$

Proposition

Under the continuity assumption, we have

$$g^* \in \arg\min_g \mathcal{R}_{\lambda^*}(g)$$

Data driven procedure: *post-processing approach*

Objective

- Estimate $g^*(x,s) \in \arg\max_k \left(p_k(x,s) - \frac{s}{\pi_s}\lambda_k^* \right)$

Different approaches

- data transformation Zemel *et al.* (2013), Calmon *et al.* (2016)
- in-processing Agarwal *et al* (2018), Donini *et al.* (2018)
- post-processing Hardt *et al.* (2016), Le Gouic *et al.* (2020)

Plug-in approach

- labeled sample $\rightarrow$ estimate p_k
- unlabeled sample $(X_1, S_1), \dots, (X_N, S_N)$
- $\{S_1, \dots, S_N\} \rightarrow$ estimate π_s by their empirical frequencies
- $\{X_1, \dots, X_N\} \rightarrow$ estimate parameter λ_k^*
- fairness guarantee requires continuity assumption

Randomization

- introduce $(\zeta_k)_{k\in\mathcal{Y}}$ i.i.d. from $\mathcal{U}_{[0,u]}$
- $\bar{p}_k(x, s, \zeta_k) = \hat{p}_k(x, s) + \zeta_k$

Randomized fair classifier

- $(X_1, \ldots, X_N) \to (X_1^s, \ldots, X_{N_s}^s)$ i.i.d. from $X|S = s$
- estimator $\hat{\lambda}$

$$\hat{\lambda} \in \arg\min_{\lambda\in\mathbb{R}^K} \sum_{s\in\mathcal{S}} \sum_{i=1}^{N_s} \max_k \left(\bar{p}_k(X_i^s, s, \zeta_{k,i}^s) - \frac{s}{\hat{\pi}_s}\lambda_k \right)$$

- resulting classifier

$$\hat{g}(x, s) \in \arg\max_{k\in\mathcal{Y}} \left(\bar{p}_k(x, s, \zeta_k) - \frac{s}{\hat{\pi}_s}\hat{\lambda}_k \right)$$

Theoretical guarantees: *fairness guarantee*

Unfairness measure

$$\mathcal{U}(g) = \max_k \left| \mathbb{P}\left(g(X,S) = k | S = 1\right) - \mathbb{P}\left(g(X,S) = k | S = -1\right) \right|$$

Distribution free-result

There exists C depending only on K and π_s such that for any estimator $\hat{p}_k$

$$\mathbb{E}\left[\mathcal{U}(\hat{g})\right] \leq C N^{-1/2}$$

Theoretical guarantees: *consistency*

Measure of performance

- $g^* \in \arg\min_g \mathcal{R}_{\lambda^*}(g)$

$$\mathcal{R}_{\lambda^*}(g) = \mathbb{P}\left(g(X,S) \neq Y\right) + \sum_{k=1}^{K} \lambda_k^* \sum_{s \in \mathcal{S}} s\mathbb{P}(g(X,S) = k | S = s)$$

- $\|\hat{\mathbf{p}} - \mathbf{p}\|_1 = \sum_{k=1}^{K} |\hat{p}_k(X,S) - p_k(X,S)|$

Theorem

Under continuity assumption

$$\mathbb{E}\left[\mathcal{R}_{\lambda^*}(\hat{g}) - \mathcal{R}_{\lambda^*}(g^*)\right] \lesssim \mathbb{E}\left[\|\hat{\mathbf{p}} - \mathbf{p}\|_1\right] + u + N^{-1/2}$$

- assume that $\hat{p}_k$ are consistent and $u \to 0$
  $\hookrightarrow$ $\hat{g}$ is consistent

Extension to ε-fairness

Approximate fairness: ε-DP

- g is ε-fair if $\mathcal{U}(g) \leq \varepsilon$

Optimal ε-fair classifier

- $g_\varepsilon^* \in \arg\min_g\{\mathbb{P}(g(X,S) \neq Y), g \text{ satisfies } \varepsilon - \text{DP}\}$
- $\left(\lambda^{*(1)}, \lambda^{*(2)}\right)$ minimizer of

$$\sum_{s \in \mathcal{S}} \mathbb{E}_{X|S=s}\left[p_k(X,s) - \frac{s}{\pi_s}\left(\lambda_k^{(1)} - \lambda_k^{(2)}\right)\right] + \varepsilon\left(\lambda_k^{(1)} + \lambda_k^{(2)}\right)$$

- $g_\varepsilon^*(x,s) \in \arg\max_k \left(p_k(x,s) - \frac{s}{\pi_s}\left(\lambda_k^{*(1)} - \lambda_k^{*(2)}\right)\right)$

Estimation

- same procedure as for exact fairness
- fairness and risk guarantees

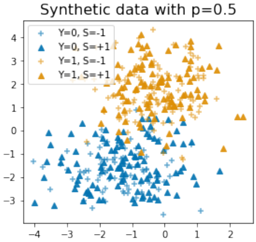
Numerical illustration: *model*

Synthetic data: *Gaussian mixture*

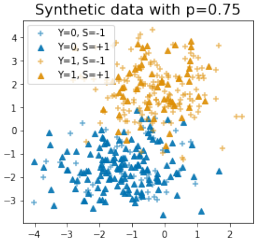
- let $c^k \sim \mathcal{U}_d(-1,1)$, and $\mu_1^k, \ldots, \mu_m^k \sim \mathcal{N}_d(0, I_d)$
- covariates: $(X|Y=k) \sim \frac{1}{m}\sum_{i=1}^{m} \mathcal{N}_d(c^k + \mu_i^k, I_d)$
- sensitive feature:

$$
\begin{aligned}
(S|Y=k) &\sim 2\cdot\mathcal{B}(p) - 1, k \leq \lfloor K/2 \rfloor \\
(S|Y=k) &\sim 2\cdot\mathcal{B}(1-p) - 1, k > \lfloor K/2 \rfloor
\end{aligned}
$$

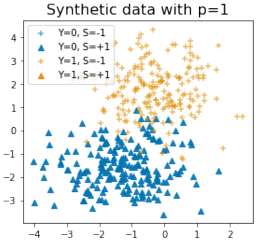
- fair data $p = 0.5$ / unfair data $p \in \{0, 1\}$



(1)



(2)

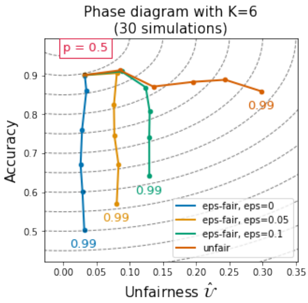


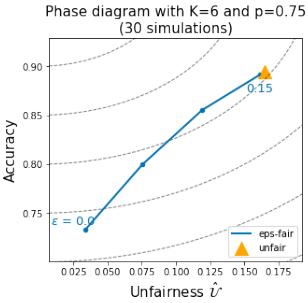
(3)

Numerical illustration: *results*

Scheme

- generate 5000 examples
- train/test/unlabeled $= 60\%/20\%/20\%$
- estimate p_k on *train* dataset using random forests
- build $\hat{g}$ using *unlabeled* dataset
- evaluated $\mathrm{Acc}(\hat{g})$ and $\mathcal{U}(\hat{g})$ using *test* dataset





Perspective

DP multiclass classification

- exact and ε-fairness
- Plug-in approach
- fairness and risk guarantee

Some extension

- Extension to prediction without sensitive attribute
- Extension to multiple sensitive attributes
- Extension to other fairness measures